一种基于改进特征加权的朴素贝叶斯分类算法 *

丁 月，汪学明†

(贵州大学 计算机科学与技术学院, 贵阳 550025)

摘 要：传统朴素贝叶分类算法没有根据特征项的不同对其重要程度进行划分，使得分类结果不准确。针对这一问题，引入 Jensen-Shannon（JS）散度，用 JS 散度来表示特征项所能提供的信息量，并针对 JS 散度存在的不足，从类别内与类别间的词频、文本频以及用变异系数修正过的逆类别频率这三个方面考虑，对 JS 散度进行调整修正，最后计算出每一特征项的权值，将权值带入到朴素贝叶斯的公式中。通过与其他算法的对比实验证明，基于 JS 散度并从词、文本、类别三方面改进后的朴素贝叶斯算法的分类效果最好。因此基于 JS 散度特征加权的朴素贝叶斯分类算法与其他分类算法相比，其分类性能有很大提高。

关键词：文本分类；朴素贝叶斯；JS 散度；词频；文本频率；类别频率

中图分类号：TP391.1   doi: 10.3969/j.issn.1001-3695.2018.07.0426

Naive Bayes classification algorithm based on feature weighting

Ding Yue, Wang Xueming†

(*College of Computer Science & Technology Guizhou University*, *Guiyang* 550025, *China*)

Abstract: The traditional Naive Bayes classification algorithm does not divide the importance degree according to the different feature items, which makes the classification result inaccurate. In order to solve this problem, this paper introduces Jensen-Shannon (JS) divergence and uses JS divergence to express the amount of information provided by the feature terms. Aiming at the deficiency of JS divergence, the paper consider from the three aspects of word frequency, text frequency and inverse category frequency corrected by coefficient of variation, the JS divergence is adjusted and corrected. The weights are introduced into the naive Bayes formula. Compared with other algorithms, it is proved that this method improves the naive Bias classification algorithm effectively. Therefore, compared with other classification algorithms, the performance of naive Bayesian classification algorithm based on JS divergence feature weighting is greatly improved.

Key words: text classification; Naive Bayes; Jensen-Shannon divergence; word frequency; document frequency ; class frequency

0 引言

当今互联网飞速发展，各类信息大规模的出现，如何在众多的信息中筛选出目标信息成了信息挖掘技术中的重要研究内容，数据挖掘中的各种文本分类算法将信息进行分组归类，提高了分类的准确性和高效性。目前，常用的分类算法有决策树分类算法、K-最邻近（KNN）分类算法、支持向量机（SVM）分类算法、朴素贝叶斯分类算法[1~4]。已有实验研究表明，在处理大规模的数据集时，KNN 算法会有很大的计算开销，而 SVM 算法的计算精度虽高但其时间开销较大，同样决策树算法的效率也会因为数据量的增大而降低。而朴素贝叶斯分类算法[5]在分类过程中的效率较为稳定，并且使用时所需要的参数方便获得，在数据有限的情况下也可计算。总体来说，算法相对简单高效，同时具有强大可靠的数学理论作为支撑，但是其分类精确度较差，分类结果仍有一定的提升空间。

目前，有两类针对传统朴素贝叶斯分类算法改进的方法[6]：第一种是放宽假设条件中的耦合程度，通过降低独立性的限制条件来提高分类精度，但是这种方法会大大增加计算量，代价过高，比如树增强朴素贝叶斯(TAN)[7]，TAN 在提高分类准确度的同时,其计算难度也大大增大；第二种是通过放大在文本分类起着重要作用的特征项的影响力，也就是给特征项赋予一个权值，这种方法即简单又能有效改善分类准确度。因此，本文提出了一种基于 JS 散度的特征加权算法，根据特征项对分类结果起到作用大小的不同，赋与不同的权值，达到对朴素贝叶斯

收稿日期：2018-07-09；**修回日期**：2018-08-06 **基金项目**：国家自然科学基金项目［2011］61163049，贵州省自然科学基金资助项目黔科合 J 字［2014］7641

作者简介：丁月（1994-），女，山西太原人，硕士研究生，主要研究方向为数据挖掘（m13403463138_1@163.com）；汪学明（1965-），通讯作者，男，安徽绩溪人，教授、博士，主要研究方向为数据挖掘、无线与移动通信、协议分析与模型检测、密码学与信息安全.

算法完成改进的目的。

1 相关研究

1.1 朴素贝叶斯分类算法

朴素贝叶斯算法是以贝叶斯算法[13]为基础，假设各特征条件相互独立的一种有效的分类算法。假设有类别集 C(c_1，c_2，c_3，…，c_n)和待分类的文本特征项 X(x_1，x_2，x_3，…，x_n)，朴素贝叶斯算法就是假设特征项在类别 c_a 与 c_b 之间相互独立的情况下，计算出特征项属于各个类别的概率 P(c_n|X)，所得最大值对应的类别就是该文本属于的类别 c_n。朴素贝叶斯分类公式如下：

$$c_{NB}=\text{argmaxP}(c_n)\prod_{m=1}^{j}P(x_m\mid c_n) \quad (1)$$

其中：P(c_n)代表的是所要分类的文本属于类别 c_n 的概率；P(x_m|c_n)所代表的是类别 c_n 中包含特征项 x_m 的概率。

$$\text{P}(c_n)=\frac{\text{属于}c_n\text{类的文本数}}{\text{训练集中的文本总数}} \quad (2)$$

$$P(x_m\mid c_n)=\frac{\text{类别}c_n\text{中包含特征项}x_m\text{的文本数量}}{\text{类别}c_n\text{中的文本总数}} \quad (3)$$

但在公式中，前提条件假设每个特征向量完全独立，并且每个特征项的权重都是相同的，不符合实际情况，得到的分类结果必然不准确。

1.2 特征加权算法

有许多专家利用属性加权赋值的算法对朴素贝叶斯分类模型进行了深入的研究和改进完善，最常用的关于计算特征权重算法的算法有词频（TF）、逆文本频（IDF）、信息增益（IG）、互信息（MI）、期望交叉熵（ECE）等。单丽莉等人[9]对 TFIDF、MI、IG、ECE 这四种常用的权重算法进行了比较研究，并提出了改进方法，用改进后的算法对旅行类的相关文本进行分类，得出文本中不存在的词语对分类起到的干扰作用比带来的贡献大的结论，因此使用 ECE 算法比用 IG 算法计算出的权重值准确。饶丽丽等人[10]对传统的 TFIDF 权重算法进行了改进，基于传统的词频与逆文本频，结合特征项在类内和类外的分布情况，提出了 TFIDF-FC 算法，并把 TFIDF-FC 算法运用到朴素贝叶斯分类算法中，分类效果得到改善； Wang 等人[11]关注到了词与类别之间的关系，提出将逆类频率（ICF）与关联频率（RF）相结合,提出若一个特征项在多数的类别中都存在，则要降低特征项的权重，通过减少特征项的权重来提高分类的准确度；针对 IG 算法没有考虑到特征项频数的问题，石慧等人[12]研究发现 IG 算法在计算特征项所提供的信息量时，只考虑了含有特征词的文本数占训练集中所有包含特征项的文本数的比例，而没有考虑特征项本身出现的次数，于是提出将类内词频与类间词频引入到传统的 IG 算法中来提高分类的准确度； Peng 等人[13]提出了一种基于相对文本分布频率的 IG 算法，用文本的相对频率分布代替了不同类别间含有特征项的文本的频率分布，进一步提高了分类性能。

虽然这些方法对提高分类的准确度起到了一定的作用，但只是在单个片面的角度进行改进，没有将特征项所携带的信息量、词频、文本频率、类别频率及其在类别内和类别外的分布情况进行综合考虑。因此，本文用 JS 散度表示特征项的信息量，并从特征词、文本、类别这三个维度的类别内外充分考虑，提出一种新的基于 JS 散度的特征加权朴素贝叶斯分类算法。

2 基于改进特征选择的加权朴素贝叶斯分类算法

传统的朴素贝叶斯算法将所有的特征项都视为同等重要，但实际上这些特征项在分类过程中起到的作用并不相同，这样会降低分类的精度，因此有必要使用文本特征选择算法对每个特征项进行加权，对其赋予一定的权值，提高分类性能。朴素贝叶斯公式改进为

$$c_{NB}=\text{argmaxP}(c_n)\prod_{m=1}^{j}P(x_m\mid c_n)\times\omega_{(m,n)} \quad (4)$$

其中：ω(m，n)是类别 c_n 中特征项 x_m 的权值；ω(m，n)是对特征项 x_m 在分类过程中产生作用大小的衡量，准确计算ω(m，n)是提高朴素贝叶斯分类准确性的关键。

2.1 JS 散度及其局限性

熵最初是物理学的专业术语，直到 1948 年，Shannon 将熵作为一种量化信息多少的数值，用到了信息论中，用信息熵表示信息的不确定度。信息增益是文本分类中特征项出现时的信息熵与不出现时信息熵之间的差值，表示了因特征词的存在而降低的不确定性的多少，即特征项所提供的信息量。

KL（kullback-leiber）散度[14]也叫交叉熵，与信息增益类似。KL 散度是指特征词在文档中存在和不存在这两种情况的距离差，这个差值用于表示该特征项带来的信息量。KL 散度与信息增益的不同点是：信息增益把同类文档中不存在某特征项时的情况也作为了一种影响分类结果的因素，而 KL 散度仅考虑特征项在文档中存时对分类产生的影响，虽然特征项的不存在也会对文本类起到作用，但是产生的干扰比带来的贡献大，所以用 KL 散度计算特征项的权重比信息增益计算得出的结果准确。KL 散度的计算公式为

$$\text{KL}(x_m)=P(c_n\mid x_m)\log\frac{P(c_n\mid x_m)}{P(c_n)} \quad (5)$$

其中：P(c_n|x_m)所表示的是含有特征项 x_m 的文本属于类别 c_n 的概率；P(c_n)表示类别 c_n 在全部的训练文本集中所占的比例。

在概率论和数理统计中，JS 散度[15]是基于散度的衡量两种概率分布相似性的一种方法,并且相对于 KL 散度有一定的优势。KL 散度具有某些局限性：a)KL 散度看似是用来表示距离的度量，其实并不具有对称性，不是真正意义上的度量；b)计算结果没有界限，不方便比较。于是本文引入 JS 散度，JS 散度是基于 KL 散度的一种变体，其不但继承了 KL 散度的优点，也弥补了上述缺陷。利用 JS 散度在计算信息熵时，其计算结果的范围始终在 0 与 1 之间，与 KL 相比，计算结果对其之间相似度的判别更加确切且方便比较。从公式中可以看出 JS 散度是具有对称性的，是真正的距离测量标准。JS 散度的计算公式如

下：

$$\mathrm{JS}(P\|Q)=\frac{1}{2}KL(P\|\frac{P+Q}{2})+\frac{1}{2}KL(Q\|\frac{P+Q}{2}) \quad (6)$$

把 JS 散度引入到朴素贝叶斯算法中，可以通过计算两种分布距离的大小来表示特征项所携带的信息量的大小，因此特征项 x_m 的 JS 熵越大，为其赋予的权值相应增大。

$$\mathrm{JS}(x_m)=\frac{1}{2}[P(c_n|x_m)\log\frac{2P(c_n|x_m)}{P(c_n|x_m)+P(c_n)}] +\frac{1}{2}[P(c_n)\log\frac{2P(c_n)}{P(c_n|x_m)+P(c_n)}] \quad (7)$$

从上述公式可以看出，JS 散度在评估一个特征项的重要性时把重心放在类别内包含特征项的文本在不同类别之间的频数和比例上，从而有三个缺点：

a)忽略特征项词频对权重的影响。若特征项在同一类别的文本中都存在，但是在每篇文本中出现的个数却很少，分布广泛但个数零星。这种情况下，该特征项的类别间文本频率高、词频低，不能很好地代表一个类别，却有着较高的权重值，这样导致了赋予的权重不准确。例如，现在有两个类别，每个类别有三篇文本，包含有两个特征项。如表 1 所示，特征项 t_1 和 t_2 都分别在类别 c_1 和 c_2 内部分布于三篇和一篇文本。根据 JS 散度计算得到的特征项 t_1 和 t_2 的权重值相等，但是特征项 t_1 在类别 c_1 内每篇文本中出现的频数都比 t_2 多，在类别 c_2 内的出现的频数却比 t_2 少，显然特征项 t_1 比 t_2 能更好的代表类别 c_1，对文本分类起到的作用更大，应该赋予更高的权重，然而并没有在 JS 散度计算公式中体现，导致误差的出现。

表 1 特征项 t1 和 t2 在不同文本中出现的频数

特征词	文本(篇)					
	C1类			C2类		
	1	2	3	1	2	3
t1	5	5	5	0	1	0
t2	1	1	1	0	3	0

b)在 JS 散度公式中，$P(c_n|x_m)$体现了含特征项的文本在类别间的聚集程度，却没有体现文本在特定类别内是否集中。如果特征项在某个类别的文本中均匀分布，则意味着这个特征项对该类别而言有一定的代表性，可以赋予这个特征项较大的权重，因此需要考虑含特征项的文本在类别内的分布比例。

c)没有考虑到类别频率对文本分类的影响，当两个特征项的词频和文本频率相同时，总的类别数与包含特征项的类别数之间的比例也是对特征项重要性判断的依据。如表 2 所示，特征项 t_3 和 t_4 在训练集中的词频和文本频率都相同，但是特征项 t_3 仅集中出现在类别 c_3 和 c_4 内，特征项 t_4 在四个类别中都有出现，含有特征项 t_3 的类别数要小于含有特征项 t_4 的类别数，特征项 t_3 分布较为集中，所以 t_3 在类别之间的区分度要大。

因此要对特征词的重要程度进行划分，就需要从词频、文本频及逆类别频率这三方面综合考虑。

表 2 特征项 t3 和 t4 在不同类别中出现的频数

特征词	文本（篇）																			
	C3类					C4类					C5类					C6类				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
t3	5	5	5	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	0	0
t4	5	5	5	0	5	5	0	0	0	0	5	0	0	0	0	5	0	0	0	0

2.2 特征项词频 TF

词频指的是特征项 x_m 存在于文本 d 内的频率。在进行文本分类的过程中，最终目的是把文本与类别进行划分和归类，而不是把文本与文本区分开来。因此在对文本分类时，首先，需要统计特征项在类别间的词频，即特征项 x_m 在类别 c_n 中出现的次数与训练集内特征项 x_m 在所有类别中出现的总数之间的比例，比值越大，说明特征项 x_m 在各类别之间分布越集中，其区分能力就越好；其次，需要统计特征项在类别内的词频，也就是特征项 x_m 在类别 c_n 各个文本中的分布情况，即特征项 x_m 在类别 c_n 中出现的次数与类别 c_n 内中文本数之比，若在类别内含有该特征项的文本数越多，分布越分散，说明该特征项的类别代表性就越好。

特征项类间词频 TBCF（term between class frequency)：

$$TBCF(x_m)=\frac{tf_n(x_m)}{\sum_{n=1}^{i}tf_i(x_m)} \quad (8)$$

特征项类内词频 TICF (term in class frequency)：

$$TICF(x_m)=\frac{tf_n(x_m)}{D} \quad (9)$$

其中：$tf_n(x_m)$表示特征项 x_m 在类别 c_n 中的频数；$tf_i(x_m)$表示特征项 x_m 在类别 c_i 中的频数；D 代表类别 c_n 中的总文本数。特征项词频的计算如下：

特征项词词频 TF(term frequency)：

$$\mathrm{TF}=TBCF\times TICF \quad (10)$$

2.3 文本频率 DF

JS 散度公式中的 $P(c_n|x_m)$项，是由含有特征项 x_m 且属于类别 c_n 的文本数除以训练集中所有含有特征项 x_m 的文本总数计算得出，表示了包含特征项的文本在类别之间的频率，因此，只需加入含有特征项文本在类别内的频率[16]即可。频率越大，表示特征项在类别内的文本中存在越普遍，说明该特征项对于该类别有很好的代表性。

文本类内频率 DICF (document in class frequency)：

$$\mathrm{DF}=\mathrm{DICF}(x_m)=\frac{df_n(x_m)}{D} \quad (11)$$

其中：$df_n(x_m)$表示含有特征项 x_m 的文本在类别 C_n 中的文本数；D 代表类别 C_n 中的总文本数。

2.4 类别频率 CF

在判定一个特征项对文本分类所起到作用的重要程度时，除了要考虑到词频和文本频率这两方面以外，类别频率也是很重要的影响因素，它能够有效的利用类间的信息。由于特定类别具有代表性的特征项似乎只存在于少数类中，所以本文使用

逆类频率来表示该项的重要水平。与逆文本频率类似，在计算逆类别频率时，特征项在类别间分布的越集中，包含特征项的类别数占总类别数的比例越小，特征项的区别度就越大。逆类别频率(ICF)：

$$\mathrm{ICF}(x_m)=\log\frac{C}{cf(x_m)} \quad (12)$$

其中：cf(x_m)表示含有特征项 x_m 的类别数；C 代表类别总数。

逆类别频率[17]中没有对类内特征项的多少进行区分，不管特征项多还是少，只要在类别中存在，就对其同等对待，这样增大了低频词的作用。因此引入类别变异系数 CCV（class coefficient of variation）度量类别频率的离散度。类别变异系数越大，证明特征项在各个类别出现频数相差越大，该特征项越具有代表性。

类别变异系数 CCV:

$$\mathrm{CCV}(x_m)=\frac{\sqrt{\dfrac{\sum_{n=1}^{C}(\mathrm{tf}_n(x_m)-\overline{\mathrm{tf}(x_m)})^2}{C}}}{\overline{\mathrm{tf}(x_m)}} \quad (13)$$

因此类别频率的计算如下：

类别频率 CF (class frequency):

$$\mathrm{CF}=ICF\times\mathrm{CCV} \quad (14)$$

2.5 改进的加权朴素贝叶斯算法

根据上述分析，本文引入 JS 散度来计算特征项的权重值，通过词频率、文本频率、类别频率对其进行修正，最终得到的特征项权值，用这个特征项权值对朴素贝叶斯公式改进。

因此最终的加权朴素贝叶斯算法如下。

$$c_{NB}=\mathrm{argmaxP}(c_n)\prod_{m=1}^{j}P(x_m\mid c_n)\times JS(x_m)\times\mathrm{TF}\times DF\times CF \quad (15)$$

3 实验研究

3.1 实验数据

为了测试本文所提出的新特征加权朴素贝叶斯分类算法的可行性和准确性，本文选取了 Sogou Labs 所提供的文本分类语料库，从中选取了汽车、IT、军事、教育、旅游、文化、体育这 7 个类别进行实验测试，每个类别选取 800 篇文本，其中 600 篇做训练文本，200 篇作为测试文本。

3.2 实验描述

实验环境如下：操作系统为 Windows 10，处理器为 Inter(R) Core i5-3210M CPU @ 2.50 GHz，内存为 4.00 GB ，开发环境是 MyEclipse 10+JDK 1.8+Tomcat 8.0，使用 Java 语言进行开发，使用中国科学院计算技术研究所开发的分词系统 NLPIR 2016 进行分词。

实验步骤主要分为分类器训练与分类器分类两个部分。实验流程如图 1 所示。具体实现步骤如下：

a)文本预处理。将训练文本使用 NLPIR 2016 实施分词、删除停用词等操作。

b)特征项选取。采用 IG 算法算出每个特征项的权值，按权值由高到低排序，对前 N 个的特征项完成挑选。

c)文本表示。把选取的特征项用向量空间模型表示，组成特征项向量组，用这个空间向量表示文本。

d)分类器训练。用本实验选取的六种改进后的朴素贝叶斯分类算法对分类文本进行分类。具体过程为：

(a)对待分类的文本进行分词、去除停用词等操作后，按 IG 算法提取一部分特征项，构成特征向量。

(b)统算出公式中的 P(cn)和 P(xm|cn)，同时根据特征加权算法算出每个特征项的权重，带入到朴素贝叶斯公式中。

(c)将分类测试文本与各个类别进行比较后，计算结果最大的就是待分类文本所属类别。

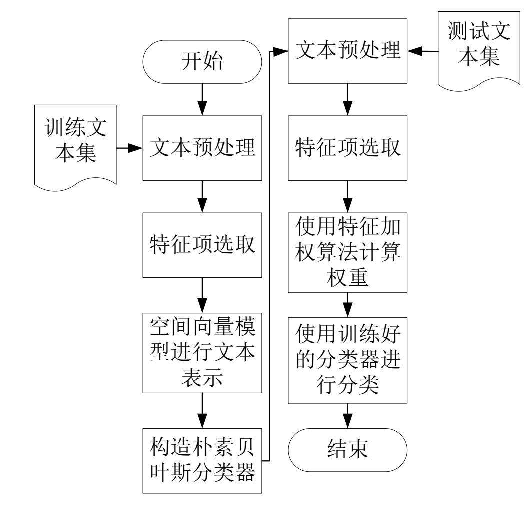


图 1 实验流程

3.3 评价方法

为了对上述各种算法的优劣性进行评估，本次实验选取精确率 P、召回率 R 和它们的调和平均值 F1 作为比较依据。

$$准确率\mathrm{P}=\frac{正确分到类别c_n中的文本数}{所有分到类别c_n中的文本数} \quad (16)$$

$$召回率\mathrm{R}=\frac{正确分到类别c_n中的文本数}{实际类别c_n中应有的文本数} \quad (17)$$

$$\mathrm{F}_1=\frac{\mathrm{P}\times R\times 2}{P+R} \quad (18)$$

3.4 实验结果

使用朴素贝叶斯（NB）算法、基于 TF 的朴素贝叶斯分类算法（TF-NB）、基于 TF-IDF 的朴素贝叶斯分类算法（TFIDF-NB）、基于 TFICF 的朴素贝叶斯分类算法（TFICF-NB）、基于 JS 散度的朴素贝叶斯分类算法（JS-NB）以及本文提出在 JS 散度的基础上进行词频、文本频率、类别频率补充修正的算法（JS-TFDFCF-NB）这六种算法进行了两组实验。

实验 1 选取特征项个数 N 的值为 500、1 000、1 500、2 000、3 000、4 000、5 000、6 000、8 000、10 000 这 10 个维度进行

实验，分析各种算法在不同维度下的分类性能。实验结果如表3和图2所示。

实验2　在特征项维度 N=3000 的条件下，比较各种算法在所选取的搜狗实验室的七个不同类别内的分类性能。实验结果如表4和图3所示。

从表3、4和图2、3中可以看出，传统朴素贝叶斯算法在六种算法中，其准确率、召回率和 F1 这三个值皆为最低，分类性能最差。其他经过加权处理后的朴素贝叶斯算法与传统朴素贝叶斯算法相比，其分类效果都有一定提高，其中由本文提出的 JS-TFDFCF-NB 算法的分类效果最好。如表3和图2所示，随着特征项维度的增多，算法分类性能随之提高，但是提高的速率越来越小，到一定程度后文本分类准确率、召回率及 F1 的值都稳定在一定范围。根据表4和图3所示，在不同类别下各分类算法的分类效果不同，在军事和体育类别分类效果较好，各个类别中 JS-TFDFCF-NB 算法比其他分类算法准确度都高。

由实验1、2的结果可以得出，用新提出的基于 JS 散度并在词频、文本频率、类别频率三方面改进的算法进行分类时，其精确度、召回率及调和平均值都相对较高，说明本文所提出的算法的分类性能相对于其他算法有了进一步的提高，是一种较好的分类算法。

表3　各算法在不同维度下实验结果统计

维度	NB（%）			TF-NB（%）			TFIDF-NB（%）			TFICF-NB（%）			JS-NB（%）			JS-TFDFCF-NB（%）		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
500	54.35	57.24	55.75	55.83	59.27	57.50	57.23	60.63	58.88	57.39	61.48	59.36	57.89	62.82	60.26	58.14	62.88	60.41
1000	61.35	65.43	63.32	65.24	69.54	67.32	65.95	64.03	64.97	60.89	65.17	62.96	66.34	70.35	68.29	69.71	73.60	71.61
1500	68.38	70.24	69.30	69.47	71.89	70.66	72.56	77.92	75.14	68.15	75.05	71.43	70.77	78.27	74.33	75.03	78.15	76.56
2000	69.46	70.96	70.20	71.62	75.34	73.43	73.89	79.24	76.47	73.33	79.24	76.17	75.82	78.97	77.37	79.23	85.53	82.26
3000	69.61	74.56	72.00	72.87	79.49	76.03	76.97	81.03	78.95	75.53	80.30	77.84	77.24	81.73	79.42	81.33	86.18	83.68
4000	71.38	77.94	74.52	74.95	76.23	75.59	78.07	82.82	80.38	77.29	80.24	78.74	76.95	77.28	77.12	80.48	84.19	82.30
5000	73.35	76.38	74.83	74.03	72.24	73.12	78.62	81.40	79.99	78.65	83.49	81.00	79.28	84.67	81.89	82.33	84.25	83.28
6000	70.23	75.30	72.68	75.29	78.24	76.74	77.19	74.29	75.71	76.55	78.56	77.54	78.77	82.47	80.58	83.55	86.29	84.90
8000	73.83	75.03	74.43	73.72	76.29	74.98	78.08	79.78	78.92	76.84	79.29	78.04	80.04	84.39	82.16	82.84	85.39	84.10
10000	72.97	78.41	75.59	72.18	74.24	73.19	78.64	80.98	79.79	77.81	78.35	78.08	81.34	85.02	83.14	83.72	87.92	85.77

表4　在 N=3000 时各算法在不同类别中实验结果统计

类别	NB（%）			TF-NB（%）			TFIDF-NB（%）			TFICF-NB（%）			JS-NB（%）			JS-TFDFCF-NB（%）		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
汽车	73.61	78.31	75.89	74.86	83.12	78.77	77.41	88.47	82.57	76.07	84.73	80.17	75.11	80.34	77.64	81.34	89.62	85.28
IT	60.95	64.93	62.88	65.33	73.75	69.29	73.71	72.28	72.99	69.21	71.23	70.21	75.92	77.55	76.73	74.98	78.84	76.86
军事	79.71	87.58	83.46	81.44	87.83	84.51	83.16	87.78	85.41	81.60	92.42	86.67	84.63	86.03	85.32	89.13	91.22	90.16
教育	68.57	76.25	72.21	75.18	82.17	78.52	78.62	80.30	79.45	72.82	78.20	75.41	80.17	83.72	81.91	82.47	88.14	85.21
旅游	61.42	67.34	64.24	65.76	71.87	68.68	68.06	71.69	69.83	69.36	71.19	70.26	67.39	75.35	71.15	75.52	77.36	76.43
文化	67.35	68.72	68.03	71.28	75.43	73.30	76.55	77.39	76.97	77.33	79.08	78.20	73.07	78.87	75.86	77.04	83.36	80.08
体育	75.63	78.78	77.17	76.25	82.23	79.13	81.25	89.32	85.09	82.35	85.25	83.77	84.36	90.25	87.21	88.84	94.71	91.68

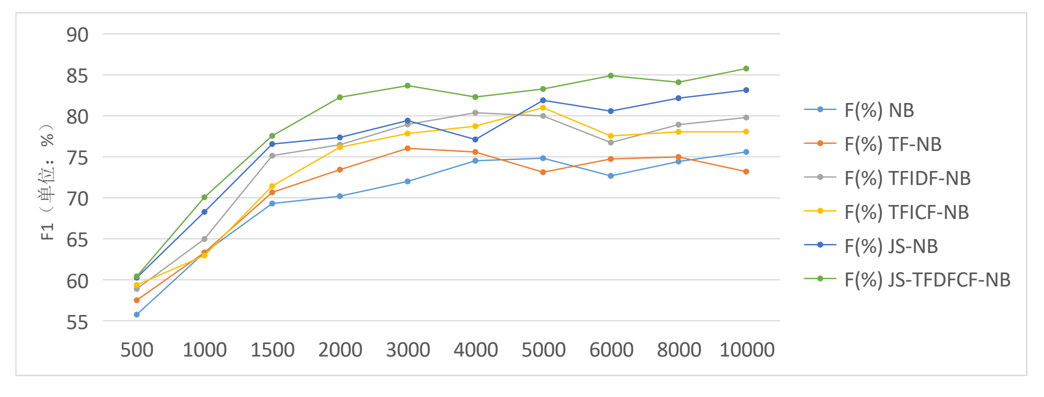


图2　各算法在不同维度下 F1 的比较

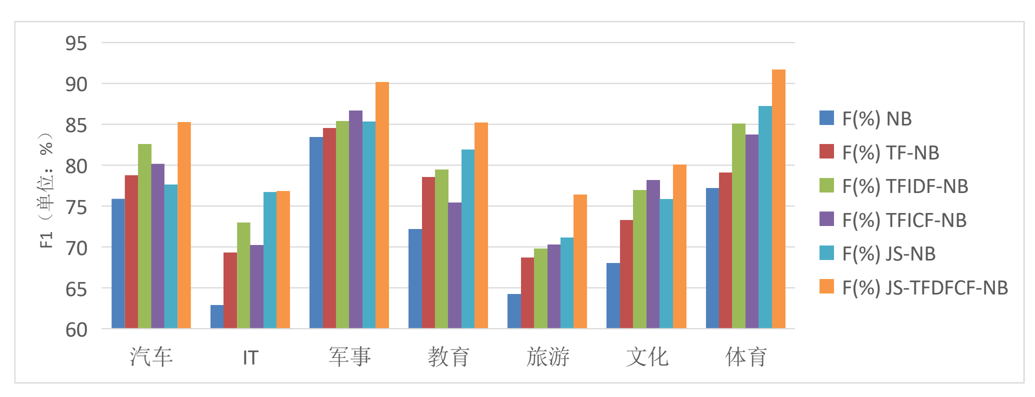


图3　各算法在 N=3000s 时在不同类别中 F1 的比较

4 结束语

在使用传统朴素贝叶斯分类算法计算分类特征项权重时，与实际情况存在出入，为完成改进，本文提出了一种新的基于特征加权的朴素贝叶斯分类算法。通过JS散度公式计算出特征项能为文本带来的信息量，然后结合类别内外的词频、文本频以及用变异系数修正过的逆类别频率对JS散度进行进一步的调整，最后使用计算出的特征项的权值改进素贝叶斯公式。通过两组对比实验证明，使用该算法可以提高朴素贝叶斯分类的准确度，并且分类效果优于其他改进算法。但是本文没有考虑特征项位置不同，其重要性也不同，接下来需要进一步的研究，继续提高分类的准确性。

参考文献：

[1] 卢苇，彭雅．几种常用文本分类算法性能比较与分析 [J]. 湖南大学学报：自然科学版，2007, 34 (6): 67-69. (Lu Wei, Peng Ya. Performance comparison and analysis of several general text classification algorithms [J]. Journal of Hunan University: Natural Sciences, 2007, 34 (6): 67-69.)

[2] Aliwy A. Comparative study of five text classification algorithms with their improvements [J]. International Journal of Applied Engineering Research, 2017, 12 (14): 4309-4319.

[3] Sohrawardi S J, Azam I, Hosain S. A comparative study of text classification algorithms on user submitted bug reports [C]// Proc of the 9th International Conference on Digital Information Management. Piscataway, NJ: IEEE Press, 2014: 242-247.

[4] Rajvanshi N, Chowdhary K R, Rajvanshi N, *et al.* Comparison of SVM and naive Bayes text classification algorithms using WEKA [J]. International Journal of Engineering & Technical Research, 2017, 6 (9): 141-143.

[5] 贺鸣，孙建军，成颖．基于朴素贝叶斯的文本分类研究综述 [J]. 情报科学，2016, 34 (7): 147-154. (He Ming, Sun Jianjun, Chengyin. Text classification based on Naive Bayes: A review [J]. Information Science, 2016, 34 (7): 147-154.)

[6] 徐光美，刘宏哲，张敬尊．基于特征加权的多关系朴素贝叶斯分类模型 [J]. 计算机科学，2014, 41 (10): 283-285. (Xu Guangmei, Liu Hongzhe, Zhang Jingzun. Multi-relational naive Bayes classifier using feature weighting [J]. Computer Science, 2014, 41 (10): 283-285.)

[7] Li Dawei, Hu Xiaojian, Jin Chengjie, *et al.* Learning to detect traffic incidents from data based on tree augmented naive Bayesian classifiers [J]. Discrete Dynamics in Nature & Society, 2017, 2017 (1): 1-9.

[8] 朱军，胡文波．贝叶斯机器学习前沿进展综述 [J]. 计算机研究与发展，2015, 52 (1): 16-26. (Zhu Jun, Hu Wenbo. Recent advances in Bayesian machine learning [J]. Journal of Computer Research and Development, 2015, 52 (1): 16-26.)

[9] 单丽莉，刘秉权，孙承杰．文本分类中特征选择方法的比较与改进 [J]. 哈尔滨工业大学学报, 2011, 43 (S1): 319-324. (Shan Lili, Liu Bingquan, Sun Chengjie. Comparison and improvement of feature selection method for text categorization [J]. Journal of Harbin Institute of Technology, 2011, 43 (S1): 319-324.)

[10] 饶丽丽，刘雄辉，张东站．基于特征相关的改进加权朴素贝叶斯分类算法 [J]. 厦门大学学报：自然科学版，2012, 51 (4): 682-685. (Rao Lili, Liu Honghui, Zhang Dongzhan. An improved weighted naive Bayes classification algorithm using feature correlation [J]. Journal of Xiamen University: Natural Science, 2012, 51 (4): 682-685.)

[11] Wang Deqing, Zhang Hui. Inverse-category-frequency based supervised term weighting scheme for text categorization [J]. Journal of Information Science & Engineering, 2013, 29 (2): 209-225.

[12] 石慧，贾代平，苗培．基于词频信息的改进信息增益文本特征选择算法 [J]. 计算机应用，2014, 34 (11): 3279-3282. (Shi Hui, Jia Daiping, Miao Pei. Improved information gain text feature selection algorithm based on word frequency information [J]. Journal of Computer Applications, 2014, 34 (11): 3279-3282.)

[13] Peng Jian, Yang Xiaohua, Ouyang Chunping, *et al.* An improved information gain algorithm based on relative document frequency distribution [C]// Proc of the 5th CCF Conference on Natural Language Processing and Chinese Computing, and the 24th International Conference on Computer Processing of Oriental Languages. New York: Springer Press, 2016: 559-567.

[14] Ren Fuji, Sohrab M G. Class-indexing-based term weighting for automatic text classification [J]. Information Sciences, 2013, 236 (1): 109-125.

[15] Lin, Jianhua. Divergence measures based on the Shannon entropy [J]. IEEE Trans on Information Theory, 1991, 37 (1): 145-151.

[16] 赵婧，邵雄凯，刘建舟，等．文本分类中一种特征选择方法研究 [J/OL]. 计算机应用研究，2019, 36 (8): 1-8. (2018-04-24) [2018-06-18]. http://www. arocmag. com/article/02-2019-08-017. html. (Zhao Jing, Shao Xiongkai, Liu Jianzhou, *et al.* Study on feature selection method in text classification [J/OL]. Application Research of Computers, 2019, 36 (8): 1-8. (2018-04-24) [2018-06-18]. http://www. arocmag. com/article/02-2019-08-017. html.)

[17] 周鹏程，刘旭敏，徐维祥．基于类别方差的特征权重算法 [J/OL]. 计算机应用研究，2018, 35 (12): 1-5. (2018-04-24) [2018-06-18]. http://www. arocmag. com/article/02-2019-08-017. html.) (Zhou Pengcheng, Liu Xuming, Xu Weixiang. Feature weighting algorithm based on class variance [J/OL]. Application Research of Computers, 2018, 35 (12): 1-5. (2018-04-24) [2018-06-18]. http://www. arocmag. com/article/02-2019-08-017. html.)